国民生活基礎調査における平均所得の推定精度の改善について

AUTHOR(S):

田栗, 正章; 井上, 隆勝

CITATION:

田栗, 正章 ...[et al]. 国民生活基礎調査における平均所得の推定精度の改善について. 数理解析研究所講究録 1989, 682: 11-37

ISSUE DATE:

1989-02

URL:

http://hdl.handle.net/2433/101165

数理解析研究所講究録
682巻 1989年 11-37

国民生活基礎調査における平均所得の推定精度の改善について

千葉大学理学部　田栗　正章（Masaaki Taguri）
千葉大学工学部　井上　隆勝（Takakatsu Inoue）

1．国民生活基礎調査の概要

厚生省では、全国の世帯及び世帯人員を対象として、「国民の基本的な生活の場である世帯の構造分析に必要な事項及び国民の保険、医療、福祉、年金、所得等、国民生活の基本事項を調査し、厚生行政の企画及び運営に必要な基礎資料を得ること」を目的として、毎年標本調査を実施している。

この調査は、昭和60年以前は厚生行政基礎調査、国民健康調査、国民生活実態調査、保険衛生基礎調査等のいくつかの調査に分かれていたが、昭和61年以降は国民生活基礎調査として一本化されて行なわれている。調査票の種類は、①世帯票、②健康票、③所得票、④貯蓄票に分かれている。そして3年に1度、大規模調査が行なわれ、中間年については、世帯票，所得票のみについて、約1/5の規模の小規模調査が行な

われている。調査対象は全国の世帯及び世帯員であり、標本抽出は国勢調査区を１次抽出単位とする、層化集落抽出法が用いられている。例えば昭和61年に行なわれた国民生活基礎調査では、昭和60年の国勢調査区約75万の中から4966調査区（約24万世帯）を抽出し、世帯票等の調査を行なった。さらにその中から、940調査区（約4万世帯）を抽出して、所得票等の調査を行なった（図１参照）。

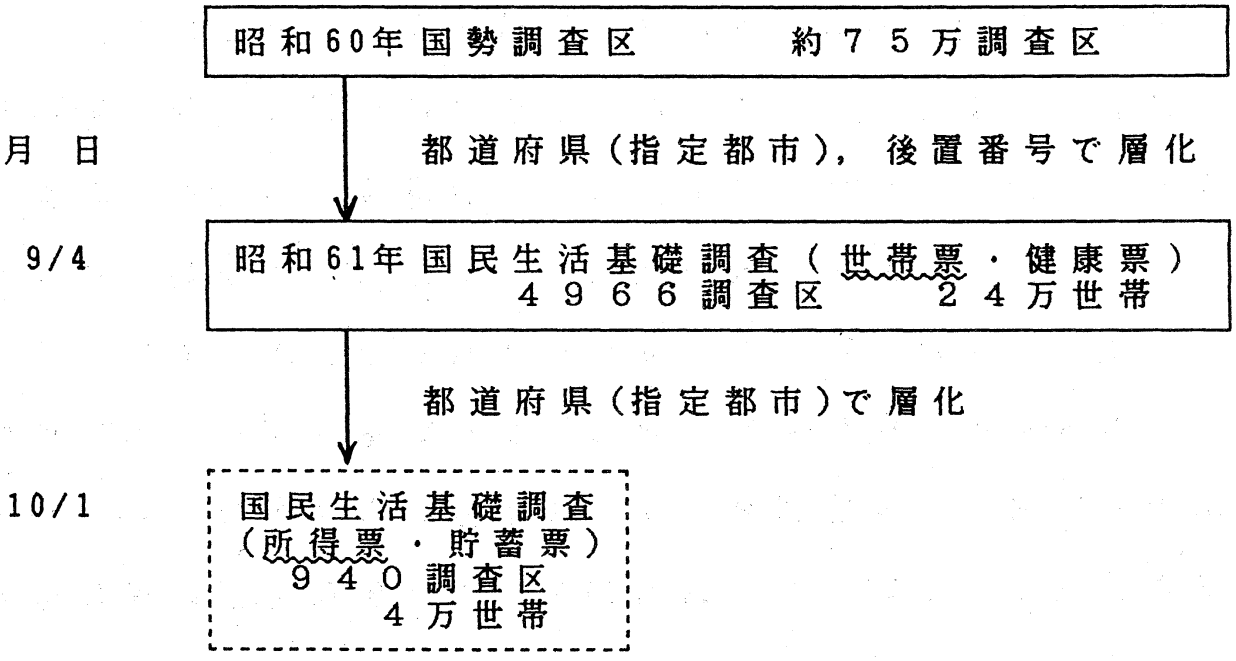

図１．国民生活基礎調査における標本抽出方法

またそれらの調査の期日は、世帯票，健康票については昭和

61年9月4日、所得票，貯蓄票については昭和61年10月1日であった。調査の方法は、世帯票，所得票については調査員の世帯訪問による面接聞き取り方式であり、健康票，貯蓄票については世帯に調査票を配布し、後日回収（貯蓄票は密封回収）とする方式であった。調査の機関と実施方法は、世帯票，健康票については厚生省→都道府県（指定都市）衛生主管部（局）→保険所→調査員という形で、また所得票，貯蓄票については厚生省→都道府県（指定都市）民生主管部（局）→福祉事務所→調査員という形で行なわれた。

2．解析の目的と方針

図1から判るように、所得票，貯蓄票に関する調査は、世帯票，健康票の調査対象の中から標本を抽出して行なわれている。しかるに現行の推計方式では、例えば世帯の平均所得の推定は、所得票の調査対象から得られるデータのみに基づいて行なわれている。しかし世帯票の調査項目の中には、世帯の平均所得と関連を持つものもあると考えられるので、所得票の調査より大規模な世帯票の調査の結果を利用することにより、推定精度が改善できる可能性がある。

そこで本研究では、このような考え方に基づいて、地域ブロック別、世帯主の年齢階級別に、世帯の平均所得の推定精

度の向上を目的とする、いくつかの推定方式を提案し、従来から使用されてきた方式も含めて、比較、検討を行なう。

上述したように、解析の方針としては、世帯の平均所得の値y（目的変数）と世帯票の各調査項目の値x（説明変数ベクトル）との間の相関関係を利用して、重回帰モデルにより平均所得推定値の精度の改善を計る（図２参照）。

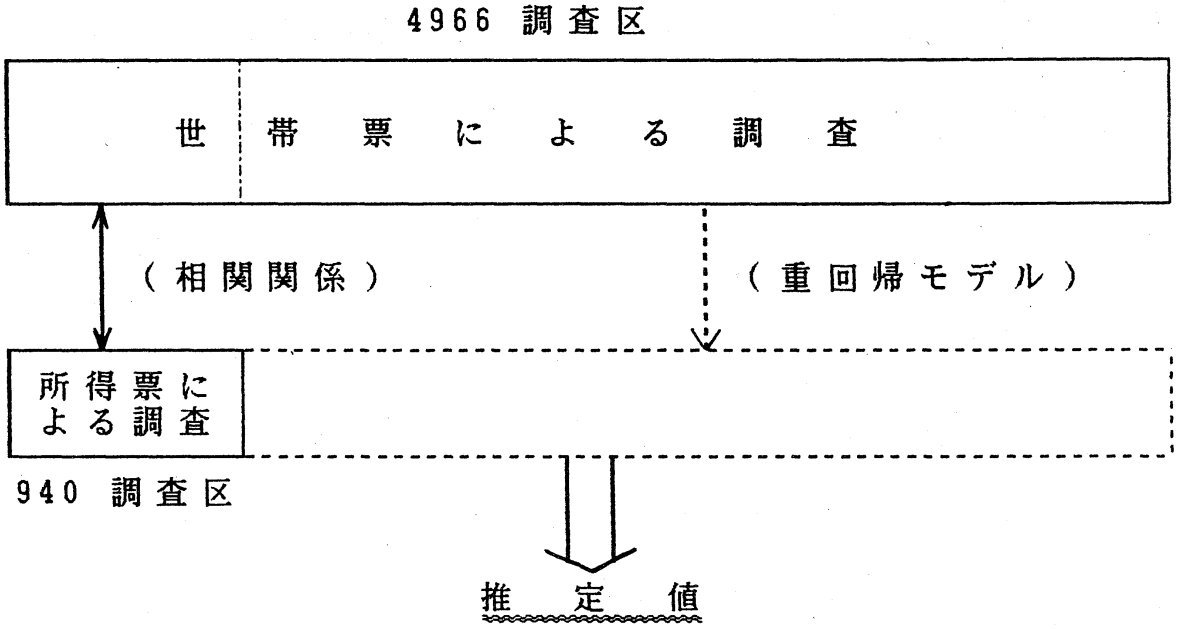

図２．提案する推定方式の模式図

３．シミュレーションと評価の方法

ある地域ブロックにおける、世帯主の年齢階級別の世帯平均所得を推定する、５種類の方法を比較、検討するために、

シミュレーション実験を行なう。使用するデータは、昭和61年の国民生活基礎調査における、世帯票および所得票による調査結果のデータである。前節で述べたように、我々の解析では1世帯の所得（総所得）の平均値を精度よく推定するために、世帯票のいくつかの項目の調査結果を用いるが、ここで使用する調査項目およびデータの型等は、表1に示す通りである。

表1．世帯所得推定のために用いる世帯票の調査項目

変数名	項目	データの型
y	総所得	実数型
x_1	世帯人員	実数型
x_2	有業人員	実数型
x_3	家計支出額	実数型
x_4	世帯主年齢	実数型
x_5	夫婦組数	実数型
d_1	世帯構造	ｶﾃｺﾞﾘｰ型
d_2	世帯業態	ｶﾃｺﾞﾘｰ型
d_3	世帯種	ｶﾃｺﾞﾘｰ型

世帯主年齢の階級の区分は、実際に使われているものと同じで、次の通りである。

階級1：　～29歳　階級2：30～39歳　階級3：40歳～49歳

階級4：50歳～59歳　階級5：60～69歳　階級6：70歳～

さて我々はいくつかの推定方法の比較を行ないたい訳であるが、そのためには母集団の情報が既知でなければならない。そこで本シミュレーション実験において想定する母集団は、ある地域ブロックにおいて、世帯票・所得票両方のデータが得られた世帯とする。ここで解析の対象とした母集団の大きさは、Ｎ＝６１５８世帯であった。標本抽出方法としては、上記の母集団からの、非復元単純無作為抽出法を用いた。また標本の大きさは、ｎ＝３００，６００等とした。現実には集落抽出法が使われているが、これは主として実際の場での調査コストの軽減を計るための抽出法であると思われる。しかも各集落内でのバラツキは集落間でのバラツキに比べて小さいと考えられるので、集落抽出法を単純無作為抽出法に置き換えてシミュレーション実験を行なったとしても、推定量の精度の比較という観点からはあまり問題はないものと考える（第７節参照）。

シミュレーションのやり方は、まず上で想定した母集団からの標本抽出実験を、Ｍ回行なう。ここで、次のように記号を定める。

(y_s, x_s)：毎回の実験により得られた標本（s＝1～n）

μ_k　　：年齢階級ｋにおける平均所得（k＝1～6）

$\hat{\mu}_k(i)$　：μ_k に対する推定値（i:推定方式、i＝1～5）

毎回の実験で得られた標本に基づいて、$\hat{\mu}_k(1)$～$\hat{\mu}_k(5)$を計算する（k＝1～6）。次にi番目の推定方法の、k番目の年齢階級について、M回のシミュレーション実験に亘る、次の量を計算し、推定方法の精度の評価を行なう（k＝1～6）。

推定値の平均　　$AV(\hat{\mu}_k) = \Sigma\ \hat{\mu}_{km} / M$

推定値の標準誤差　　$VAR(\hat{\mu}_k) = \Sigma\ \{\hat{\mu}_{km} - AV(\hat{\mu}_k)\}^2 / M$

推定値のM.S.E.　　$MSE(\hat{\mu}_k) = \Sigma\ \{\hat{\mu}_{km} - \mu_k\}^2 / M$

以上のシミュレーション実験の概略を図３に示す。

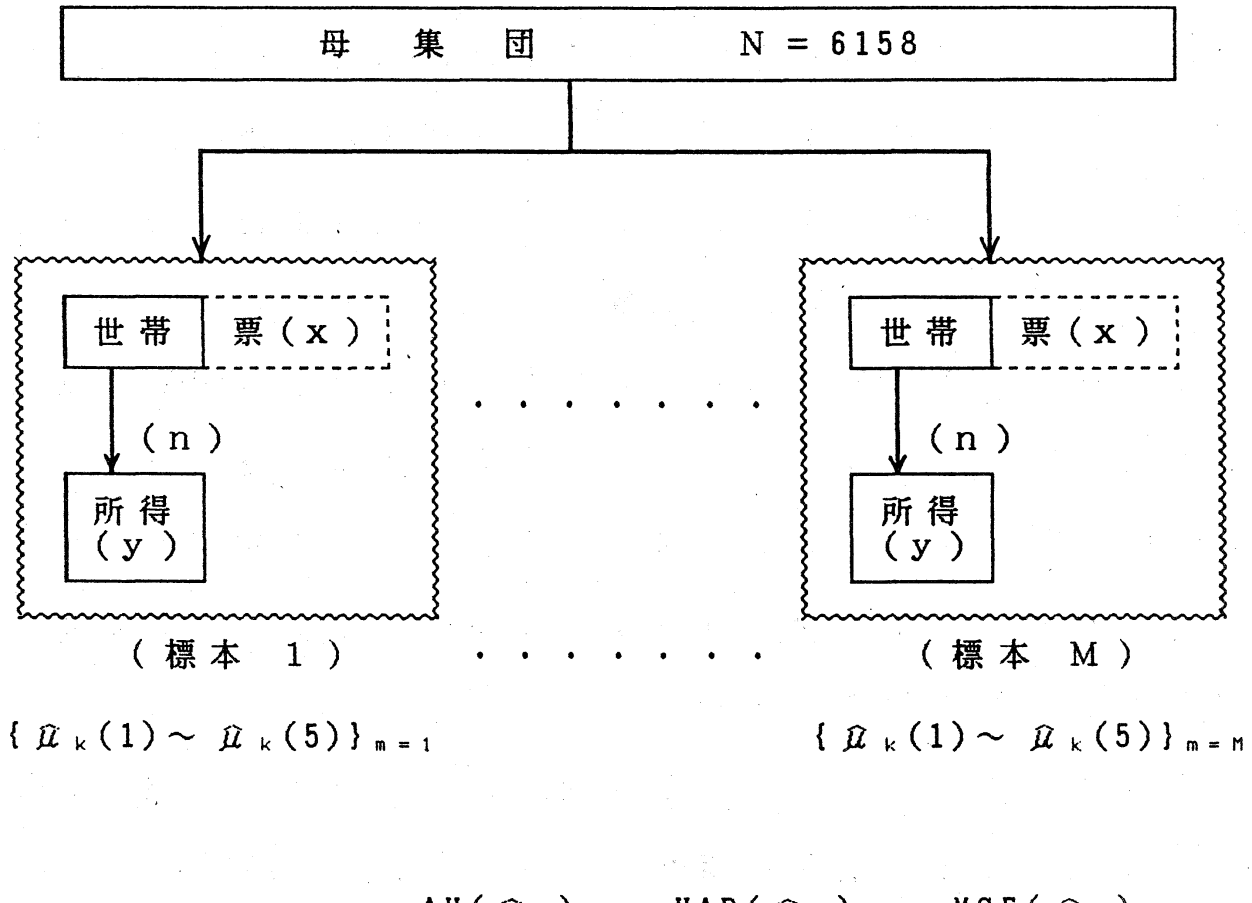

図３．シミュレーション実験の模式図

4．予備的なデータ解析

推定目標であるμ_kを推定するために、次の重回帰モデルを考える。

$$y = x'\beta + \varepsilon, \qquad \varepsilon \sim (0, \sigma^2)$$

ここでεは平均0,分散σ^2をもつある分布に従うとする。μ_kの推定値としては、

$$\hat{\mu}_k = \bar{x}_k' \bar{\beta}$$

を用いる。ここで$\bar{x}_k$はk番目の年齢階級における説明変数ベクトルxの平均で、与えられるものとする。また$\bar{\beta}$は、次節で述べるいくつかの方法によって推定されたβの値とする。

ここで与えられたデータの特性を把握するために、いくつかの予備的なデータ解析を行なう。まず上の重回帰モデルで用いる目的変数および説明変数について検討する。この場合、目的変数yは世帯の総所得であり、これは実数型変数である。また実数型の説明変数xとしては、表1に示したように世帯人員，有業人員，家計支出額，世帯主年齢，夫婦組数を考える。これに対してカテゴリー型の説明変数dは、世帯構造，世帯業態，世帯種である。

まず上の5つの実数型説明変数のみを用いて重回帰分析を行なったところ、重相関係数Rの値は0.5より小さく、精度の良い推定を行なうことは難しいと考えられた。そこで変数変

換や合成変数の作成等を行なうことにした。さらにデータを層別する必要もあると判断した。何種類かの変数変換や合成変数の作成を試行錯誤で行なった結果、本解析では目的変数としてはlog(y)またはyを用いることにした。また合成変数としては $x_6=(x_3)^{1/4}$ [x_3は家計支出額] および $x_7=(x_1*x_3)^{1/4}$ [x_1は世帯人員] を追加することにした。合成変数x_7は、1人当りの所得(総所得／世帯人員)が家計支出額に比例するとして導き出されたものである。これらの変数を加えた7つの説明変数を用いた重回帰分析では、重相関係数Rの値は約0.6となった。

次に、カテゴリー型の説明変数について検討を行なった。これらの変数のカテゴリーは、次の通りである。

世帯構造 : 1.単独 2.夫婦のみ 3.夫婦と子
4.片親と子 5.三世代 6.その他

世帯業態 : 1.役員 2.小企業 3.大企業 4.日雇
5.自営雇人有 6.自営雇人無 7.内職
8.専業農家 9.兼業農家

世帯種 : 1.国民保険 2.被雇用者保険 3.両方

上記の 6 x 9 x 3 個の各カテゴリー(層)における度数, yの平均, 標準偏差, yとxとの散布図, 相関係数の値等を求め、これらの結果を参考にしてカテゴリー(層)の合併を行

い、ダミー変数を決定することにした。

ここではyとxとの散布図の1例として、総所得yと家計支出額x_3およびそれに関連する合成変数x_6, x_7との関係を図4に与えておく。例えばこの図を見ると、x_3そのものより合成変数x_6, x_7を考えることにより、yとの相関が高まることが判る。

次に3つのカテゴリー型説明変数の各々について、カテゴリー別に世帯の総所得の平均値を求めたところ、表2のようであった。さらにこれらのカテゴリーに関する同種の3次元のクロス表も作成し、これらの結果を参考にして表2に示したようなカテゴリーの合併（層の構成）を行なうことにした。すなわち層の数は、全部で 2 x 3 x 2 個とした。次節で考える重回帰モデルでは、これらの層に対応する 12－1＝11 個（計画行列のランク落ちを防ぐため11とする）のダミー変数を用いる。

5．重回帰モデルと5種類の推定方式

本節では、ここで提案する4種類の推定方式と、従来からの方式について考える。前節で述べたように、考える重回帰モデルは

$$y = x'\beta + \varepsilon, \qquad \varepsilon \sim (0, \sigma^2)$$

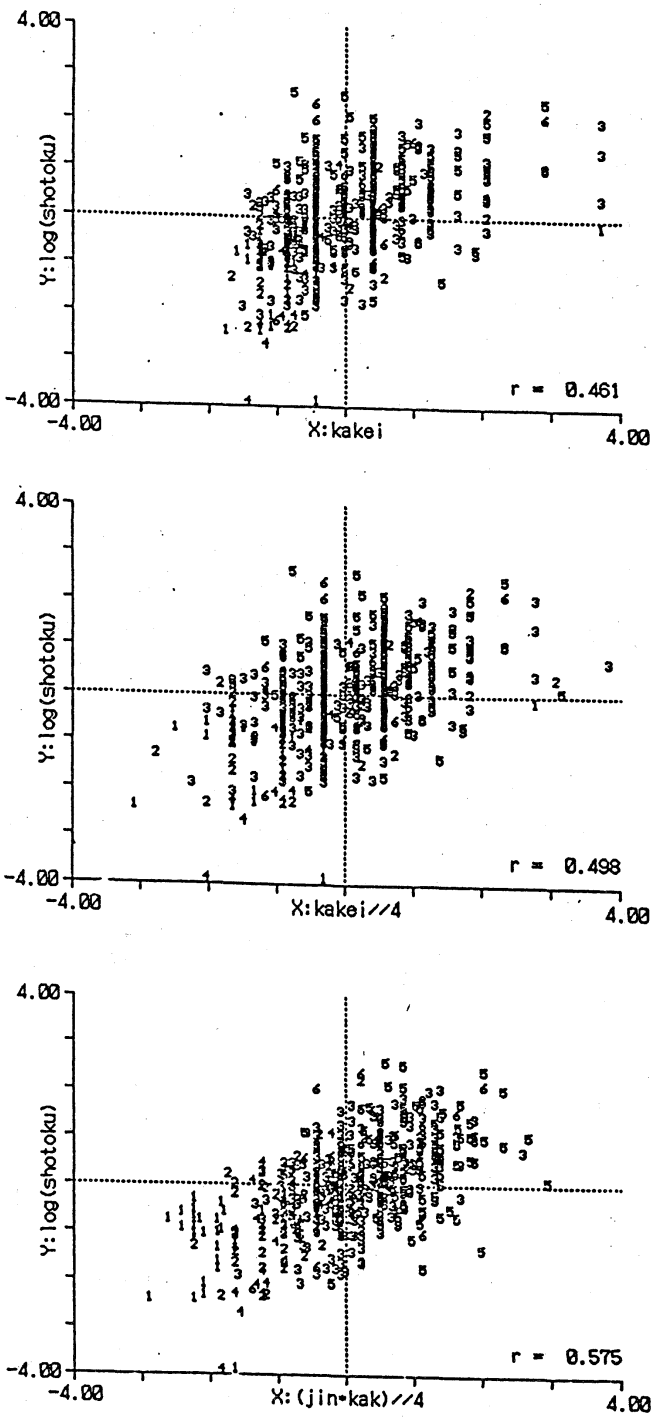

図４．総所得（ｙ）と、家計支出額に関連した３種類の変数（x_3，x_6，x_7）の散布図　［世帯構造別］

表２．カテゴリー別に見たｙの平均と層の合併

	カテゴリー	ｙの平均	層の合併
世帯構造	1．単独	5.423	
	2．夫婦のみ	6.087	層１
	3．夫婦と子	6.315	
	4．片親と子	5.872	
	5．三世代	6.433	層２
	6．その他	6.251	
世帯業態	1．役員	6.745	
	2．小企業	6.105	
	3．大企業	6.413	層２
	4．日雇	5.668	
	5．自営雇人有	6.352	層１
	6．自営雇人無	6.072	
	7．内職	5.603	層３
	8．専業農家	5.828	
	9．兼業農家	6.330	
世帯種	1．国民保険	5.881	層１
	2．被雇用者保険	6.324	層２
	3．両方	6.349	

であり、ｋ番目の年齢階級における母集団の平均所得μ_kの推定は

$$\hat{\mu}_k = \bar{x}_k' \hat{\beta}$$

によって行なう。ここで$\hat{\beta}$は、以下で述べる４種類の推定方法で求める。また$\bar{x}_k$は、ｋ番目の年齢階級の説明変数ベクトルｘの平均（推定点）で、推定を行なう場合には与えられていなければならない。実際の場では$\bar{x}_k$の値は、例えば世帯票の調査結果等から推定すべきであるが、ここではシミュレー

ション実験を簡単にするため、その第1近似として母集団(N=6158)に亘る平均を用いることにする（第7節参照）。ここで上の重回帰モデルにおける目的変数と説明変数をまとめておくと、次のようになる。

目的変数　　：　y，　log(y)　　[y：総所得]

実数型説明変数：　x_1(世帯人員)，x_2(有業人員)，

x_3(家計支出額)，x_4(世帯主年齢)，x_5(夫婦組数)，

x_6(=$\sqrt[4]{x_3}$；合成変数)，x_7(=$\sqrt[4]{x_1 * x_3}$；合成変数)

ダミー変数　　：　x_8 ～ x_{18}（11個；表2参照）

[2層(世帯構造)×3層(世帯業態)×2層(世帯種)－1＝11]

さて、このようなモデルに基づいてμ_kの推定を行なう訳であるが、推定値の安定性という観点からは、変数選択（モデル選択）を行なうべきである。ここでは実数型説明変数のみを変数選択の対象とし、ダミー変数は変数選択の対象としない強制変数とする。

ここで5種類の推定方式について説明を行なう。上述の重回帰モデルを、目的変数ベクトルy，計画行列X，回帰係数ベクトルβ，誤差ベクトルεを用いて書き直しておくと、

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim (0, \sigma^2 I_n)$$

となる。ここでI_nはn次の単位行列である。推定目標は

$$\mu_k = \bar{x}_k' \beta$$

であり、一般にK個の点の上での推定を考える（k = 1, 2, ... , K）。モデル(変数)選択の対象とする各モデルを識別するモデル添え字ベクトルをi(p)、対応するモデル M[i(p)] を $y = X_i \beta_i + \varepsilon$ で表す。また、モデル M[i(p)] の下での推定量を $\hat{\mu}_k = \bar{x}_{ik}' \bar{\beta}_i$ と書く。

次に、モデル選択の規準とするリスク関数 $r(\hat{\mu}_k)$ について考える。我々の問題では、各推定点上での推定量の精度を同等に評価すべきであると考えられた。すなわち、どの年齢階級における推定量のバラツキも同等に評価すべきであると考えられたので、$r(\hat{\mu}_k)$ としては、次の量を用いる。

$$r(\hat{\mu}_k) = K^{-1} \Sigma \; E\{(\mu_k - \bar{x}_{ik}' \bar{\beta}_i)^2\}$$

ここで、$\Delta_{ii} = K^{-1} \Sigma (\bar{x}_{ik} \bar{x}_{ik}')$ とおき、Δ_{ii}^{+} を Δ_{ii} の Moore-Penrose型一般逆行列で定義すると、モデル M[i(p)] の下でリスク関数 $r(\hat{\mu}_k)$ を最小にする回帰係数ベクトル $\beta_{i\cdot}$ は、$\beta_{i\cdot} = \Delta_{ii}^{+} \Delta_{iq} \beta$ で表せる。また、リスク関数は次式のように分解される。

$$\begin{aligned} r(\hat{\mu}_k) &= K^{-1} \Sigma \; E\{[(\bar{x}_k' \beta - \bar{x}_k' \beta_{i\cdot}) \\ &\qquad + (\bar{x}_k' \beta_{i\cdot} - \bar{x}_{ik}' \bar{\beta}_i)]^2\} \\ &= (\beta' \Delta_{qq} \beta - \beta_{i\cdot}' \Delta_{qq} \beta_{i\cdot}) \\ &\qquad + E\{(\beta_{i\cdot} - \bar{\beta}_i)' \Delta_{ii} (\beta_{i\cdot} - \bar{\beta}_i)\} \\ &= M.B.(\beta_{i\cdot}) + MSE_{\Delta}(\bar{\beta}_i) \end{aligned}$$

ここで、右辺第２式の第１項 M.B.($\beta_{i\cdot}$)は、最大モデル(全ての説明変数を取り込んだモデル)のかわりにモデルM[i(p)]を用いたために生じる Model Bias項と考えられる。同第２項 $MSE_{\Delta}(\bar{\beta}_i)$は、モデルM[i(p)]の下での推定点$x_k$に対する積和行列$\Delta_{ii}$を計量とした場合の、$\bar{\beta}_i$の平均２乗誤差(MSE)と考えられるが、これはさらに次のように分割できる。

$$MSE_{\Delta}(\bar{\beta}_i) = (\beta_{i\cdot} - E\{\bar{\beta}_i\})'\Delta_{ii}(\beta_{i\cdot} - E\{\bar{\beta}_i\}) + tr[\Delta_{ii}V(\bar{\beta}_i)]$$

ここで右辺第１項はΔ_{ii}を計量とした$\bar{\beta}_i$の$\beta_{i\cdot}$に対する推定Biasに、また第２項は推定分散である。すなわち、モデルM[i(p)]の下での推定量 $\hat{\mu}_k = x_{ik}'\bar{\beta}_i$ のリスク r($\hat{\mu}_k$)は、Model Bias項, $\bar{\beta}_i$の推定Bias項, 推定分散項に分解できる。したがって、モデルM[i(p)]及びβ_iの推定量$\bar{\beta}_i$の選び方により、異なる推定方式を考えることができる。まず$\bar{\beta}_i$に関しては、ここでは次の２種類を考える。

$$\bar{\beta}_i : \begin{cases} \hat{\beta}_i = (X_i'X_i)^{-1}(X_i'X)\beta, \Rightarrow & E\{\hat{\beta}_i\} \neq \beta_{i\cdot} \\ \tilde{\beta}_i = \Delta_{ii}^{+}\Delta_{iq}\beta, \Rightarrow & E\{\tilde{\beta}_i\} = \beta_{i\cdot} \end{cases}$$

推定量$\hat{\beta}_i$はモデルM[i(p)]の下での通常の最小２乗(OLS)推定量 $(X_i'X_i)^{-1}X_i'y$ であるが、$\beta_{i\cdot}$に対する不偏推定量とはならない。一方、$\tilde{\beta}_i$は推定域(K個の推定点)上での修正最小２乗(Modified-OLS)推定量で、$\beta_{i\cdot}$の不偏推定量となる。

次にモデル選択のための規準統計量に着目する。まず$\Delta_{qq} = n^{-1}X'X$とした場合、すなわち推定域と標本域における積和行列が一致した場合には、以下の各式が成立する。

$$\beta_{i\cdot} = (n^{-1}X_i'X_i)^{-1}(n^{-1}X_i'X)\beta = E\{\hat{\beta}_i\} \equiv \beta_i(X)$$

$$M.B.(\beta_{i\cdot}) = \beta'(n^{-1}X'X)\beta - \beta_i(X)'(n^{-1}X_i'X_i)\beta_i(X)$$

$$MSE_{\Delta}(\hat{\beta}_i) = p\sigma^2/n$$

したがって、この場合のリスク関数の1つの不偏推定量は、MallowsによるCp規準と一致する。しかし、推定域が標本域と一致しない我々の場合には、推定域を考慮したリスク関数$r(\hat{\mu}_k)$に基づいてモデル選択を行う必要がある。表3の2、3の推定方式では$\beta_{i\cdot}$の推定量として$\hat{\beta}_i$及び$\tilde{\beta}_i$を用いたときのリスク関数値の不偏推定量を各々構成し、それらを規準統計量としてモデル選択を行った。これらの組合せの他に、ここでは毎回のモデル選択を行なう際に、推定量$\hat{\beta}_i$, $\tilde{\beta}_i$の選択を伴ったモデル選択方式（SEL.と略す）も考える。したがって提案する推定量は4種類となる。以下ではこれらの推定量に、従来から用いられてきた年齢階級別単純平均$\bar{y}$（Y-SMPと略す）を加えた5種類の方式について検討を行なうが、それらの記号と簡単な説明を表3にまとめておく。

表３．用いた推定方式

#	記号	β_iの推定量	選択規準
1.	OLS-X :	$\hat{\beta}_i = (X_i'X_i)^{-1}(X_i'X)$	標本域での評価
2.	OLS :	$\hat{\beta}_i = (X_i'X_i)^{-1}(X_i'X)$	推定域での評価
3.	M-OLS :	$\tilde{\beta}_i = \Delta_{ii}^{+}\Delta_{iq}\hat{\beta}$	推定域での評価
4.	SEL. :	$\hat{\beta}_i$, $\tilde{\beta}_i$ の選択	推定域での評価
(5.)	Y-SMP :	$\bar{y}$（階級別単純平均）	（従来からの方法）

６．推定結果

前節で与えた５種類の推定方式を用いて、第３節で述べたようなシミュレーション実験を、各ケースについてＭ＝100回ずつ行なった。ここではその内の主な計算結果を表として与え、検討を行なう。

まず５種類の方式についての比較を行なうと、例えば表４のようになる。ここで各年齢階級の番号は、第３節で与えたものである。またμ_kはここで想定した母集団でのｋ番目の年齢階級における母平均であり、REG_kは$\bar{x}_k$に対する母集団の回帰値である。平均$AV(\bar{\mu}_k)$について考えると、いずれの方式に

よる推定値もμ_kに近いが、Y-SMPが比較的良さそうである。ここで注目すべきことは、μ_kがk=5（5番目の年齢階級）で一度小さくなってから、k=6でまた大きくなっているのに対応して、いずれの推定値でも全く同じ傾向になっていることである。Y-SMPはμ_kの不偏推定量であるので当然であろうが、提案した方式は重回帰モデルに基づいているので、母集団の構造をうまく反映していないとこのような傾向は見られないであろう。この意味では、我々のモデルは現実の場での使用に一応耐え得るものと考える。

次に平均2乗誤差MSE($\hat{\mu}_k$)については、ここで提案した4種類の方式は、従来からの方法（Y-SMP）と比べて約1/6程度に改善されていることが判る。また4種類の方式の中では、M-OLSが良さそうである。そこで以下では、主としてM-OLSとY-SMPについての検討を行なう。以上述べた傾向は、他の場合にも共通して観察された。したがってこの実験に関する限りでは、ここで提案した推定方式は、従来から用いられてきた方法を大きく改善していると考えられる。中でも推定域を考慮して推定を行なう方式は、改善の効果が高いと考えられる。従来の方法では、年齢階級の上の方で平均2乗誤差が極端に大きくなっているが、例えばM-OLSはそのようなことがないように作られているので、全体としての平均2乗誤差も小さく

なっていると考えられる。

次に、目的変数の対数変換の影響について検討すると、表5および表6のようになる。表5より平均AV($\hat{\mu}_k$)に関しては、yとlogyとの差はあまり認められない。しかし表6を見ると、MSE($\hat{\mu}_k$)に関してはyとlogyの場合で大きな差が観察される。すなわちlogyを目的変数とした場合には、提案した方式の精度は従来の方法とほぼ同程度であるが、yの場合には上述したように約1/6に改善されている。この傾向は他のケースについても共通して観察されたので、以下では目的変数としてはyを用いることにする。この場合以下の表7に示すように、重相関係数の値は、最大モデルの下でも、R＝0.4程度であり、個々のデータの回帰平面のまわりのバラツキはかなり大きい。しかし推定点（各年齢階級の平均値）における提案した推定値の、母数（真の値）のまわりのバラツキは、従来からの方法に比べてかなり小さい。

目的変数がyの場合、母集団におけるyとxとの回帰構造がどのようになっているかを見るために、この場合の重回帰分析等の結果を表7に与える。ここで注目すべきことは、各年齢階級における各変数の平均値$\bar{x}_{kj}$（jは変数の番号）と目的変数y（$=\mu_k$）との相関が0.8～0.9と極めて高いことである（表7の一番下の表を参照のこと）。したがってこの関係を

係を用いれば、μ_kの非常に良い推定値が得られるのではないかと考えられる。しかしこれは母集団における$\bar{x}_{kj}$とμ_kとの相関であり、標本からμ_kを推定する場合には、母集団における関係が近似的にでも成立しているか否かは判らないので、各年齢階級の平均値のみを用いた回帰分析を行なった。その結果の１例が、表８である。回帰に用いた説明変数はこの場合３つであり、表中に示されている。y(=μ_k)と$\bar{x}_k$との重相関係数はほぼ1であり、この意味では回帰は非常にうまくいっている。他の場合も解析したが、ほぼ同様の結果であった。そこで各推定量$\hat{\mu}_k$のシミュレーションに亘る平均AV($\hat{\mu}_k$)と平均２乗誤差MSE($\hat{\mu}_k$)を計算したところ、表９のようになった。AV($\hat{\mu}_k$)に関しては、μ_kはk=4のところで最大となっており、Y-SMPの場合も同様であるが、提案した方式の場合は$\hat{\mu}_k$の値はkに関して単調増加となっており、この意味で回帰はうまくいっていない。またMSE($\hat{\mu}_k$)についても、例えばM-OLSとY-SMPとの差はあまり大きくない。これはk=1とkの大きな階級における平均２乗誤差が、かなり大きくなっているためである。このように、母集団の構造としては各変数の平均値$\bar{x}_{kj}$とμ_kはほぼ線形な関係をもっているが、標本から推定点上における推定を行なうと、精度は良くない。

最後に、標本数ｎの変化に対する平均２乗誤差MSE($\hat{\mu}_k$)へ

の影響について考える。n = 300, 600, 1200の場合の結果を示すと、表10のようになる。nが小さくなるとMSE($\hat{\mu}_k$)の値は大きくなっているが、大きくなる割合はどの方式でもほぼ一定のように見える。したがって推定方式の比較だけを行なう場合には、例えばn = 600の場合について考えておけばよいであろう。

7．結果のまとめと今後の課題

ここでは、前節で検討した事柄を簡単にまとめておく。

(1) ここで提案した、重回帰モデルを用いる推定方式による推定値の精度は、従来からの推定値（各年齢階級における単純平均）の精度に比べて高い。特に n = 300（抽出率1/20）の場合には、平均2乗誤差は、1/6程度になる。

(2) 提案した推定量の内では、推定域を考慮した Modified OLS推定量による推定方式を用いたものが、平均2乗誤差の意味では最も良さそうである。したがって推定域を考慮した重回帰分析を行なう必要がある。

今後の検討課題としては、次のような問題が考えられる。

① 現実により近い状況の下での検討

ここで行なったシミュレーション実験では、標本抽出は単純無作為抽出法を用いたが、現実には集落抽出法が使われて

いる。したがって現実と同じ集落抽出法の場合について、検討を行なうことが必要である。また本実験では、推定点としては想定した母集団の平均値$\bar{X}_k$を用いたが（第5節参照）、現実の場では例えば世帯票の情報等から$\bar{X}_k$を推定しなければならない。この推定が最終的に推定量の精度に及ぼす影響についても検討しておくことが必要である。

② 提案した手法の安定性の検討

ここでは1つの地域ブロックのみについての検討を行なったが、提案した方法を他の地域ブロックに対して適用し、結果の安定性について十分な検討を行なうべきである。他の地域ブロックについて、上と同様な解析を試みた結果では、$N \fallingdotseq 2500$, $n \fallingdotseq 300$のとき（抽出率$\fallingdotseq 1/8$）、Modified OLS推定量を用いた推定値の平均2乗誤差は、従来の方法に比べて約2/3程度になった。

③ 重回帰モデルの改善

残差の検討を行い、目的変数も含めてより良い合成変数，変数変換等を捜し、回帰の精度を上げることが必要である。また層別方法を再検討し、それに応じて適当なダミー変数を導入することも必要であろう。

表４．５種類の推定量の比較［ｎ＝600，目的変数：ｙ の場合］

(1) 各方式の下での推定値のシミュレーションに亘る平均AV($\bar{\mu}_k$)

年齢階級k	1:OLS-X	2:OLS	3:M-OLS	4:SEL.	Y-SMP	μ_K	REG_K
1	301.	300.	296.	297.	309.	303.	299.
2	531.	531.	532.	532.	533.	532.	531.
3	638.	638.	636.	637.	624.	625.	637.
4	704.	705.	705.	705.	718.	718.	708.
5	681.	680.	683.	682.	695.	690.	680.
6	765.	764.	763.	763.	754.	751.	761.

(2) 各方式の下での推定値のシミュレーションに亘る平均２乗誤差MSE($\bar{\mu}_k$)

年齢階級k	1:OLS-X	2:OLS	3:M-OLS	4:SEL.	Y-SMP
平均	1530.	1500.	1380.	1410.	8360.
1	1160.	1200.	1090.	1080.	599.
2	418.	422.	416.	400.	442.
3	555.	561.	526.	537.	801.
4	1200.	1190.	1150.	1170.	1890.
5	1660.	1670.	1580.	1590.	5780.
6	4180.	3960.	3490.	3650.	40700.

表５．目的変数の対数変換に対する検討［ｎ＝600］
（各方式の下での推定値のシミュレーションに亘る平均）

(1) ｙ の場合

年齢階級k	1:OLS-X	2:OLS	3:M-OLS	4:SEL.	Y-SMP	μ_K	REG_K
1	301.	300.	296.	297.	309.	303.	299.
2	531.	531.	532.	532.	533.	532.	531.
3	638.	638.	636.	637.	624.	625.	637.
4	704.	705.	705.	705.	718.	718.	708.
5	681.	680.	683.	682.	695.	690.	680.
6	765.	764.	763.	763.	754.	751.	761.

(2) log(ｙ) の場合

年齢階級k	1:OLS-X	2:OLS	3:M-OLS	4:SEL.	Y-SMP	μ_K	REG_K
1	5.65	5.65	5.65	5.65	5.56	5.55	5.65
2	6.15	6.15	6.15	6.15	6.18	6.18	6.15
3	6.29	6.29	6.29	6.29	6.30	6.31	6.29
4	6.37	6.36	6.37	6.37	6.39	6.39	6.37
5	6.29	6.28	6.29	6.29	6.30	6.27	6.28
6	6.41	6.40	6.41	6.41	6.27	6.26	6.40

表６．目的変数の対数変換に対する検討［ｎ＝600］
（各方式の下での推定値のシミュレーションに亘る平均２乗誤差）

(1) ｙ の場合

年齢階級ｋ	1:OLS-X	2:OLS	3:M-OLS	4:SEL.	Y-SMP
平均	1530.	1500.	1380.	1410.	8360.
1	1160.	1200.	1090.	1080.	599.
2	418.	422.	416.	400.	442.
3	555.	561.	526.	537.	801.
4	1200.	1190.	1150.	1170.	1890.
5	1660.	1670.	1580.	1590.	5780.
6	4180.	3960.	3490.	3650.	40700.

(2) log(ｙ) の場合

年齢階級ｋ	1:OLS-X	2:OLS	3:M-OLS	4:SEL.	Y-SMP
平均	.00738	.00742	.00704	.00711	.00765
1	.01300	.01380	.01190	.01220	.00569
2	.00132	.00125	.00129	.00131	.00108
3	.00047	.00046	.00047	.00047	.00133
4	.00134	.00157	.00130	.00134	.00271
5	.00204	.00245	.00185	.00191	.00866
6	.02610	.02490	.02540	.02540	.02640

表７．母集団における重回帰分析［目的変数：ｙ］

標本数	： 6158			
目的変数	： 総所得（ｙ）			
重相関係数	: R	0.442		
寄与率	: R**2	0.196		
自由度調整済Ｒ	: RA	0.193		
残差の標準誤差	: ROOT-VE	430.3		
系列相関係数	: S-COR.	1:0.057	2:0.045	3:0.022

変数名	＋＋ 回帰係数 ＋＋ STD.B	B	D(B)	T=B/D(B)
x_1（世帯人員）	-1.731D-01	-5.834D+01	1.977D+01	-2.950D+00
x_2（有業人員）	1.516D-01	8.151D+01	7.280D+00	1.119D+01
x_3（家計支出額）	-1.009D-01	-3.027D+00	7.775D-01	-3.894D+00
x_4（世帯主年齢）	1.179D-01	4.472D+00	4.894D-01	9.138D+00
x_5（夫婦の組数）	2.788D-02	2.932D+01	1.770D+01	1.655D+00
x_6（$\sqrt[4]{x_3}$）	4.497D-02	7.985D+01	1.360D+02	5.869D-01
x_7（$\sqrt[4]{x_1 * x_3}$）	3.855D-01	3.242D+02	9.614D+01	3.372D+00
x_8（Dummy Var.）	3.033D-02	1.432D+02	7.032D+01	2.037D+00
x_9（Dummy Var.）	7.415D-02	1.494D+02	5.020D+01	2.977D+00
x_{10}（Dummy Var.）	-3.880D-02	-9.951D+01	5.356D+01	-1.857D+00
x_{11}（Dummy Var.）	1.097D-02	2.032D+01	5.000D+01	4.063D-01
x_{12}（Dummy Var.）	-2.017D-02	-6.189D+01	5.708D+01	-1.084D+00
x_{13}（Dummy Var.）	5.899D-04	2.511D+00	6.697D+01	3.749D-02
x_{14}（Dummy Var.）	3.850D-02	7.753D+01	4.800D+01	1.615D+00
x_{15}（Dummy Var.）	1.160D-01	1.170D+02	4.352D+01	2.689D+00
x_{16}（Dummy Var.）	-7.887D-02	-1.290D+02	4.603D+01	-2.803D+00
x_{17}（Dummy Var.）	-3.316D-03	-3.705D+00	4.389D+01	-8.442D-02
x_{18}（Dummy Var.）	-6.227D-02	-1.904D+02	5.477D+01	-3.477D+00
定数項	0.0	-6.555D+02	1.172D+02	-5.589D+00

表７．母集団における重回帰分析［目的変数：ｙ］（つづき）

＋＋ 分散分析表 ＋＋

要因	自由度	平方和	平均平方	F
回帰	18	2.76569D+08	1.53650D+07	8.29829D+01
残差	6139	1.13669D+09	1.85158D+05	
合計	6157	1.41326D+09		

＝＝ 各年齢階級におけるｙの平均 ＝＝

k	階級	度数	ｙの平均	回帰平面上の値
1:	-29	0.092	303.	299.
2:	30-39	0.271	532.	531.
3:	40-49	0.279	625.	637.
4:	50-59	0.227	718.	708.
5:	60-69	0.089	690.	680.
6:	70-	0.042	751.	761.

＊＊ 目的変数，実数型説明変数の平均，標準偏差，相関行列 ＊＊

変数番号	平均	S.D.	1	2	3	4	5	6	7	y
x_1	0.36D+01	0.14D+01	1.0	0.4	0.2	0.2	0.7	0.4	0.8	0.2
x_2	0.16D+01	0.89D+00	0.4	1.0	0.1	0.4	0.4	0.2	0.3	0.3
x_3	0.26D+02	0.16D+02	0.2	0.1	1.0	0.1	0.2	0.9	0.7	0.2
x_4	0.45D+02	0.13D+02	0.2	0.4	0.1	1.0	0.3	0.2	0.2	0.2
x_5	0.93D+00	0.46D+00	0.7	0.4	0.2	0.3	1.0	0.3	0.6	0.2
x_6	0.22D+01	0.27D+00	0.4	0.2	0.9	0.2	0.3	1.0	0.9	0.3
x_7	0.30D+01	0.57D+00	0.8	0.3	0.7	0.2	0.6	0.9	1.0	0.3
y	0.60D+03	0.48D+03	0.2	0.3	0.2	0.2	0.2	0.3	0.3	1.0

＊＊目的変数，実数型説明変数の平均，標準偏差，相関行列＊＊
（各年齢階級の平均値に対する）

変数番号	平均	S.D.	1	2	3	4	5	6	7	y
x_1	0.35D+01	0.81D+00	1.0	0.7	0.9	0.8	1.0	0.9	1.0	0.9
x_2	0.16D+01	0.42D+00	0.7	1.0	0.7	0.9	0.8	0.7	0.8	0.9
x_3	0.25D+02	0.40D+01	0.9	0.7	1.0	0.5	0.8	1.0	1.0	0.8
x_4	0.50D+02	0.19D+02	0.8	0.9	0.5	1.0	0.9	0.5	0.7	0.9
x_5	0.93D+00	0.26D+00	1.0	0.8	0.8	0.9	1.0	0.8	0.9	1.0
x_6	0.22D+01	0.11D+00	0.9	0.7	1.0	0.5	0.8	1.0	1.0	0.8
x_7	0.29D+01	0.33D+00	1.0	0.8	1.0	0.7	0.9	1.0	1.0	0.9
y	0.60D+03	0.17D+03	0.9	0.9	0.8	0.9	1.0	0.8	0.9	1.0

表８．各年齢階級の平均値のみを用いた回帰分析
（母集団における重回帰分析［目的変数：ｙ］）

標本数	:	97
目的変数	:	総所得（ｙ）
重相関係数	: R	0.974
寄与率	: R**2	0.948
自由度調整済Ｒ	: RA	0.946
残差の標準誤差	: ROOT-VE	27.99
系列相関係数	: S-COR.	1:0.903 2:0.806 3:0.710

＋＋ 分散分析表 ＋＋

要因	自由度	平方和	平均平方	F
回帰	3	1.32483D+06	4.41611D+05	5.63758D+02
残差	93	7.28501D+04	7.83334D+02	
合計	96	1.39768D+06		

＋＋ 回帰係数 ＋＋

変数名	STD.B	B	D(B)	T=B/D(B)
x_1(世帯主年齢)	6.374D-01	6.259D+00	2.856D-01	2.191D+01
x_2(家計支出額)	2.603D-01	9.471D+00	2.766D+00	3.423D+00
x_3($\sqrt[4]{}$世人*家計)	2.168D-01	1.067D+02	3.947D+01	2.704D+00
定数項	0.0	-2.476D+02	5.221D+01	-4.742D+00

＝＝ 各年齢階級おけるｙの平均 ＝＝

k	階級	度数	ｙの平均	回帰平面上の値
1:	-29	0.092	303.	320.
2:	30-39	0.271	532.	522.
3:	40-49	0.279	625.	649.
4:	50-59	0.227	718.	677.
5:	60-69	0.089	690.	707.
6:	70-	0.042	751.	804.

＊＊ 目的変数，実数型説明変数の平均，標準偏差，相関行列 ＊＊

変数番号	平均	S.D.	1	2	3	y
x_4	0.45D+02	0.12D+02	1.0	0.5	0.6	0.9
x_3	0.26D+02	0.33D+01	0.5	1.0	0.9	0.8
x_7	0.30D+01	0.25D+00	0.6	0.9	1.0	0.8
y	0.60D+03	0.12D+03	0.9	0.8	0.8	1.0

＊＊目的変数，実数型説明変数の平均，標準偏差，相関行列＊＊
（各年齢階級の平均値に対する）

変数番号	平均	S.D.	1	2	3	y
x_4	0.50D+02	0.19D+02	1.0	0.5	0.7	0.9
x_3	0.25D+02	0.40D+01	0.5	1.0	1.0	0.8
x_7	0.29D+01	0.33D+00	0.7	1.0	1.0	0.9
y	0.60D+03	0.17D+03	0.9	0.8	0.9	1.0

表９．各年齢階級の平均値のみを用いた回帰分析
（５種類の推定量の比較［ｎ＝600，目的変数：ｙ の場合］）

(1) 各方式の下での推定値のシミュレーションに亘る平均

年齢階級ｋ	1:OLS-X	2:OLS	3:M-OLS	4:SEL.	Y-SMP	μ_k	REG_k
1	320.	324.	320.	320.	309.	303.	320.
2	523.	522.	520.	520.	533.	532.	522.
3	646.	648.	649.	649.	624.	625.	649.
4	675.	677.	678.	678.	718.	718.	677.
5	709.	708.	709.	709.	695.	690.	707.
6	808.	805.	806.	806.	754.	751.	804.

(2) 各方式の下での推定値のシミュレーションに亘る平均２乗誤差

年齢階級ｋ	1:OLS-X	2:OLS	3:M-OLS	4:SEL.	Y-SMP
平均	4340.	3990.	4080.	4080.	8360.
1	2630.	1930.	2280.	2260.	599.
2	843.	736.	930.	929.	442.
3	1720.	1550.	1520.	1520.	801.
4	2880.	2790.	2670.	2670.	1890.
5	4760.	4600.	4580.	4580.	5780.
6	13200.	12300.	12500.	12500.	40700.

表10．標本数の変化に対する平均２乗誤差の検討［目的変数：ｙ の場合］

(1) ｎ＝３００ の場合

年齢階級ｋ	1:OLS-X	2:OLS	3:M-OLS	4:SEL.	Y-SMP
平均	3050.	3030.	2580.	2830.	14800.
1	2170.	2110.	1570.	1790.	1410.
2	957.	937.	744.	789.	884.
3	1200.	1190.	1070.	1120.	2140.
4	2330.	2380.	2420.	2430.	3230.
5	3780.	3800.	3210.	3520.	10700.
6	7860.	7780.	6440.	7310.	70400.

(2) ｎ＝６００ の場合

年齢階級ｋ	1:OLS-X	2:OLS	3:M-OLS	4:SEL.	Y-SMP
平均	1530.	1500.	1380.	1410.	8360.
1	1160.	1200.	1090.	1080.	599.
2	418.	422.	416.	400.	442.
3	555.	561.	526.	537.	801.
4	1200.	1190.	1150.	1170.	1890.
5	1660.	1670.	1580.	1590.	5780.
6	4180.	3960.	3490.	3650.	40700.

(3) ｎ＝１２００ の場合

年齢階級ｋ	1:OLS-X	2:OLS	3:M-OLS	4:SEL.	Y-SMP
平均	565.	564.	546.	547.	3210.
1	392.	386.	398.	397.	290.
2	112.	111.	118.	118.	172.
3	238.	247.	224.	228.	344.
4	604.	633.	604.	606.	692.
5	696.	702.	668.	670.	2890.
6	1350.	1300.	1270.	1260.	14900.